Use R!

Christian Ritz
Jens Carl Streibig

Nonlinear Regression with R

Springer

Christian Ritz
Department of Basic Sciences
and Environment (Statistics)
Faculty of Life Sciences
University of Copenhagen
Thorvaldsensvej 40
DK-1871 Frederiksberg C
Denmark
ritz@life.ku.dk

Jens Carl Streibig
Department of Agriculture and Ecology
(Crop Science)
Faculty of Life Sciences
University of Copenhagen
Hoejbakkegaard Allé 13
DK-2630 Taastrup
Denmark
jcs@life.ku.dk

ISBN: 978-0-387-09615-5 e-ISBN: 978-0-387-09616-2
DOI: 10.1007/978-0-387-09616-2

Library of Congress Control Number: 2008938643

Printed on acid-free paper

springer.com

model. Typically only one or a few – if any – of these hypotheses are relevant to consider. Let us have a look at the summary output for the model fit `secalonic.m1`.

```
> summary(secalonic.m1)

Formula: rootl ~ SSfpl(dose, a, b, c, d)

Parameters:
  Estimate Std. Error t value Pr(>|t|)
a 6.053612   0.395467  15.308 0.000606 ***
b 0.353944   0.194089   1.824 0.165722
c 0.075188   0.005911  12.721 0.001048 **
d 0.029350   0.006621   4.433 0.021333 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2028 on 3 degrees of freedom

Number of iterations to convergence: 0
Achieved convergence tolerance: 8.101e-06
```

The parameters a and b are the upper and lower limits on the observed root length. The logistic model will predict root lengths close to a for doses almost 0, and the predicted root lengths will approach b as the dose becomes very large. It would be natural to expect that the lower limit will be close to 0, as large doses may prohibit any growth, and therefore it is relevant to consider the hypothesis $H_0 : b = 0$. The value of the corresponding t-test statistic is 1.82. The corresponding p-value based on the t-distribution with 3 degrees of freedom is 0.166. This means that we cannot reject the hypothesis $H_0 : b = 0$, tentatively suggesting that the lower limit could be 0. The result is not surprising in light of Fig. 7.4. Consequently, we can simplify the original four-parameter logistic model by fixing the lower limit at 0 to obtain a three-parameter model describing the data. This three-parameter model is also available in a self-starter version for `nls()`, namely `SSlogis()` (we will use it in the next subsection).

A related approach for testing the hypothesis $H_0 : \beta_j = \beta_{j0}$ would be to use confidence intervals: If β_{j0} is not contained in the 95% confidence interval of β_j, then the hypothesis in Equation (7.3) is rejected at a significance level of 5%. So this approach can be applied using either profile, bootstrap, or Wald confidence intervals.

7.5.2 Using F-tests

The t-tests rely on the estimated standard errors and thus rely to a large extent on the linear approximation of the mean function. By using the F-test,

we can curb the influence on the linear approximation. In order to apply the F-test, two models need to be fitted: Model A and Model B. Model B should be a submodel of Model A; that is, obtained from Model A by imposing some constraint on the parameters. The choice of submodel will be determined by the null hypothesis that is of interest. The F-test statistic is defined as

$$F = \frac{(\mathrm{RSS_B}(\hat{\beta}_B) - \mathrm{RSS_A}(\hat{\beta}_A))/(df_\mathrm{B} - df_\mathrm{A})}{\mathrm{RSS_A}/df_\mathrm{A}}$$

where subscripts A and B refer to Model A and Model B, respectively. This test statistic is related to the statistic introduced in Subsection 5.2.3. The test is sometimes referred to as the extra-sum-of-squares F-test (Motulsky and Christopoulos, 2004, Chapter 22). The main ingredient in the statistic is the difference between the RSS quantities for the two models considered. A large difference means that the two models are quite different, whereas a small difference indicates that they provide similar fits to the data. Large and small can be quantified by means of a p-value obtained from an F-distribution with degrees of freedom $(df_\mathrm{B} - df_\mathrm{A}, df_\mathrm{A})$.

Getting back to the dataset `secalonic`, the first step is to fit the submodel that we want to compare to the initial model. Thus we fit the three-parameter logistic model where the lower limit is fixed at 0. The reason for considering this submodel is of that we want to test the hypothesis that the lower limit could be 0 (just as it was in the previous subsection). As already mentioned, the relevant self-starter function is `SSlogis()`.

```
> secalonic.m2 <- nls(rootl ~ SSlogis(dose,
+     a, c, d), data = secalonic)
```

To assess whether or not the reduction from the model fit `secalonic.m1` to the model fit `secalonic.m2` is warranted by the data, the `anova` method is invoked to calculate the F-test defined above.

```
> anova(secalonic.m2, secalonic.m1)

Analysis of Variance Table

Model 1: rootl ~ SSlogis(dose, a, c, d)
Model 2: rootl ~ SSfpl(dose, a, b, c, d)
  Res.Df Res.Sum Sq Df  Sum Sq F value
1      4    0.25924
2      3    0.12341  1 0.13582  3.3016
  Pr(>F)
1
2 0.1668
```

The order of the arguments `secalonic.m2` and `secalonic.m1` does not have any effect on the resulting p-value, but we still always specify the fitted submodel as the first argument in `anova`. The F-test yields the same conclusion

as was already established using the t-test, in fact giving almost the same p-value as the t-test in the previous subsection (see Exercise 7.4 for a situation where results differ).

7.6 Non-nested models

Consider a collection of candidate models that has been chosen a priori for a particular dataset. These models need not be submodels of each other, and therefore it may not be possible to use the F-test procedure introduced in the previous section to compare these models. How then do we decide which model is the most appropriate?

Ideally, the experimenter collecting or generating the data should decide which model to use based on subject matter. If no such suggestions are available, then it may be useful to use some kind of statistic to compare the available models. The basic idea is to calculate the value of the statistic for each candidate model and then compare these values to determine which model provides the best fit. The decision rule is simple: One model is better than another model if it has the smallest value of the statistic. Based on all pairwise comparisons using this rule, a ranking of the candidate models can be established.

The residual standard error and Akaike's information criterion (AIC) (Burnham and Anderson, 2002, pp. 94–96) are two statistics, that are often used for model comparison and selection. The residual standard error is a measure of the distance between the data and fitted regression curve based on the model fit, whereas we can think of the AIC as being an estimate of the distance from the model fit to the true but unknown model that generated the data. The AIC is defined as -2 times the maximum value of the log likelihood plus 2 times the number of parameters in the model. Using Equation (2.4) in Subsection 2.2.2, the AIC can be written as

$$\begin{aligned} \text{AIC} &= -2\log\left(L(\hat{\beta},\hat{\sigma}^2)\right) + 2(p+1) \\ &= n\log(2\pi) + n\log(\text{RSS}(\hat{\beta})/n) + n + 2(p+1) \end{aligned} \tag{7.5}$$

For nonlinear regression models as defined by Equation (1.3), the AIC is a function of the residual sum of squares, the number of observations and number of parameters (Burnham and Anderson, 2002, p. 94), so the same ingredients are used in both the residual standard error and the AIC, but, as we shall see shortly, they will not in general produce the same ranking of the models. By definition, the AIC includes a penalty for the number of parameters used (the term $2(p+1)$ in Equation (7.5)). Therefore, using the AIC will take the model complexity into account, and this feature may curb overfitting.

We consider the dataset `M.merluccius`, which contains stock and recruitment data for hake in the period 1982–1996 (see also Section 1.1). Cadima (2003, pp. 47–49) considers four stock-recruitment models for these data: